A Probabilistic Formulation of Keyword Spotting

Resumen

Joan Puigcerver

November 19, 2018

La detección de palabras clave (*Keyword Spotting*, en inglés), aplicada a documentos de texto manuscrito, tiene como objetivo recuperar los documentos, o partes de ellos, que sean relevantes para una cierta consulta (*query*, en inglés), indicada por el usuario, entre una gran colección de documentos. La temática ha recogido un gran interés en los últimos 20 años entre investigadores en Reconocimiento de Formas (*Pattern Recognition*), así como bibliotecas y archivos digitales.

Esta tesis, en primer lugar, define el objetivo de la detección de palabras clave a partir de una perspectiva basada en la Teoría de la Decisión y una formulación probabilística adecuada. Más concretamente, la detección de palabras clave se presenta como un caso particular de Recuperación de la Información (*Information Retrieval*), donde el contenido de los documentos es desconocido, pero puede ser modelado mediante una distribución de probabilidad. Además, la tesis también demuestra que, bajo las distribuciones de probabilidad correctas, el marco de trabajo desarrollada conduce a la solución óptima del problema, según múltiples medidas de evaluación utilizadas tradicionalmente en el campo.

Más tarde, se utilizan distintos modelos estadísticos para representar las distribuciones necesarias: Redes Neuronales Recurrentes o Modelos Ocultos de Markov. Los parámetros de estos son estimados a partir de datos de entrenamiento, y las respectivas distribuciones son representadas mediante Transductores de Estados Finitos con Pesos (*Weighted Finite State Transducers*).

Con el objetivo de hacer que el marco de trabajo sea práctico en grandes colecciones de documentos, se presentan distintos algoritmos para construir índices de palabras a partir de modelos probabilísticos, basados tanto en un léxico cerrado como abierto. Estos índices son muy similares a los utilizados por los motores de búsqueda tradicionales.

Además, se estudia la relación que hay entre la formulación probabilística presentada y otros métodos de gran influencia en el campo de la detección de palabras clave, destacando cuáles son las limitaciones de los segundos.

Finalmente, todas la aportaciones se evalúan de forma experimental, no sólo utilizando pruebas académicas estándar, sino también en colecciones con decenas de miles de páginas provenientes de manuscritos históricos. Los resultados muestran que el marco de trabajo presentado permite construir sistemas de detección de palabras clave muy rápidos y precisos, con una sólida base teórica.